Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

Simultaneous discovery of cancer subtypes and subtype features by molecular data integration

Thanh Le Van [1], Matthijs van Leeuwen [2], Ana Carolina Fierro [3], Dries De Maeyer [4], Jimmy Van den Eynden [5], Lieven Verbeke [3], Luc De Raedt [1], Kathleen Marchal [3,4,6,7,*], Siegfried Nijssen [1,2,*]

[1]Department of Computer Science, KULeuven, Belgium,
[2]Leiden Institute for Advanced Computer Science, Universiteit Leiden, The Netherlands,
[3]Department of Information Technology, iMinds, Ghent University, Belgium,
[4]Bioinformatics Institute Ghent, Technologiepark 927, 9052 Gent, Belgium,
[5]Department of Medical Biochemisty and Cell Biology, Institute of Biomedicine, University of Gothenburg, Sweden,
[6]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium,
[7]Department of Genetics, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Subtyping cancer is key to an improved and more personalized prognosis/treatment. The increasing availability of tumor related molecular data provides the opportunity to identify molecular subtypes in a data-driven way. Molecular subtypes are defined as groups of samples that have a similar molecular mechanism at the origin of the carcinogenesis. The molecular mechanisms are reflected by subtype-specific mutational and expression features. Data-driven subtyping is a complex problem as subtyping and identifying the molecular mechanisms that drive carcinogenesis are confounded problems. Many current integrative subtyping methods use global mutational and/or expression tumor profiles to group tumor samples in subtypes but do not explicitly extract the subtype-specific features. We therefore present a method that solves both tasks of subtyping and identification of subtype-specific features simultaneously. Hereto our method integrates mutational and expression data while taking into account the clonal properties of carcinogenesis. Key to our method is a formalisation of the problem as a *rank matrix factorisation* of *ranked data* that approaches the subtyping problem as *multi-view bi-clustering*.
Results: We introduce a novel integrative framework to identify subtypes by combining mutational and expression features. The incomparable measurement data is integrated by transformation into *ranked data* and subtypes are defined as *multi-view bi-clusters*. We formalise the model using *rank matrix factorisation*, resulting in the SRF algorithm. Experiments on simulated data and the TCGA breast cancer data demonstrate that SRF is able to capture subtle differences that existing methods may miss.
Contact: kathleen.marchal@intec.ugent.be, siegfried.nijssen@cs.kuleuven.be

1 Introduction

As cancer is a heterogeneous disease, subtyping cancer is key to an improved and more personalised prognosis and treatment. With cancer genomes, transcriptomes, and epigenomes becoming increasingly available, one of the major challenges in cancer research is to use these molecular data to define clinically or biologically meaningful subtypes.

Successful seminal research on cancer subtyping aimed at grouping patients based on similarities in their molecular profiles (gene expression) or extracting expression derived features to optimally classify patients

according to clinically relevant phenotypes (Perou *et al.*, 2000; Sørlie *et al.*, 2001; Mischel *et al.*, 2003; Tothill *et al.*, 2008).

With the availability of NGS data, charting cancer genomes, transcriptomes and even epigenomes offers the opportunity to refine subtyping by taking into account not only the molecular phenotypes (expression) but also likely driver events (mutations, CNVs, methylations) that are at the origin of the tumorigenesis (Vogelstein *et al.*, 2013).

Several efforts have been taken to integrate these different molecular data in order to extract relevant subtypes, for instance Yuan *et al.* (2011) and Curtis *et al.* (2012) relied mainly on combining copy number and expression data to define subtypes, whereas the more generic models of Mo *et al.* (2013), Wang *et al.* (2014) and Speicher and Pfeifer (2015) use next to expression and CNV also mutation and methylation data.

The problem of these early approaches, which aim at clustering samples based on shared CNV and mutational profiles, is that they overlook one of the major properties of tumorigenesis: its clonality. By directly using copy number alterations to discriminate between samples they ignore the fact that CNVs are prevalent in cancerous cells and that many CNVs are passenger events, not involved in driving the phenotype ($> 70\%$) (Zack *et al.*, 2013; Sanchez-Garcia *et al.*, 2014). Using passenger events to group patients might blur the true sample grouping in the data as driving events are rare compared to passenger events.

The same goes for the sample grouping based on shared somatic mutational profiles. Doing this implicitly assumes that true driving somatic mutations are frequent across tumor samples, which is because of the clonality of the carcinogenesis not necessarily true. Because they evolve independently, tumors can trigger the same driver pathways through mutations in different genes. By focusing only on frequent alterations, rare events that are very characteristic for a subtype are ignored. In addition, if similarities between tumor samples are scored using the raw mutation data, results are mainly driven by the dense data, such as copy number and gene expression with a negligible contribution from the mutation data.

The most advanced state-of-the-art integrative methods for cancer analysis do take into account the clonal properties of cancer by searching for mutational consistency at pathway level rather than at the individual gene level. They do so by exploiting the connectivity of mutations occurring across different tumor samples on an interaction network. An interaction network here consists of a comprehensive compilation of all molecular interaction information, available on an organism of interest; the network is represented as a graph in which the nodes correspond to genes and the edges to interactions between the genes. Mutations that are recurrently affecting sets of genes that are closely connected on the interaction network are identified as drivers (Leiserson *et al.*, 2014; Verbeke *et al.*, 2015; De Maeyer *et al.*, 2016). Hofree *et al.* (2013) successfully applied this strategy to use mutation data for subtyping.

Here we introduce a novel analysis framework that combines CNVs and mutation data with an expression phenotype to identify subtypes while considering mutational consistency at a pathway level. Because identifying subtypes and defining the molecular mechanisms (driver pathways) that drive cancer are confounded (a subtype depends on the molecular mechanism but the molecular mechanisms that one can identify also depends on how patients are grouped), our method performs the two tasks simultaneously. The distinguishing feature of our method is that it is based on *ranked matrix factorisation*: all data is represented in ranked form, in which we identify factors that are used to define subtypes. We propose new methods to cast the different types of data into a ranked form. We extensively tested the performance of our method on simulated data. Comparing our method with other state-of-the-arts on the well-studied TCGA breast cancer dataset shows how our method is able to grasp the most prominent signatures in the data that are also retrieved by other methods, but also how it is able to capture subtle differences that are missed by methods that compare samples based on global profiles of similarities.

2 The SRF algorithm

An overview is given in Figure 1, details follow. Key processing steps include 1) diffusing mutation information over an interaction network on a per sample basis; 2) removing scale differences by applying a rank-based transformation of the mutation and expression data; 3) applying a model based on *rank matrix factorisation* (Le Van *et al.*, 2015) to jointly factorise the transformed data into a number of *ranked factors*. Each resulting factor consists of a subset of samples associated to a subset of expressed and mutated genes; 4) defining subtypes as combination of ranked factors.

2.1 Transforming input datasets into rank matrices

The first step is to transform the original data into rank matrices.

Transforming transcription data Given a gene expression matrix $A \in \mathbb{R}^{l \times n}$, where l is the number of expression genes and n is the number of tumor samples, its corresponding rank matrix is obtained by sorting each row's values from low to high and assigning ranks accordingly (with the largest rank being assigned to the highest value). That is, all samples are ranked for each gene. The resulting *ranked expression matrix* E has the same size as A and the values in each row are a subset of $\sigma_1 = \{1, \ldots, n\}$ (ties all get the lowest rank). Figure 1C shows an example of an expression matrix A and Figure 1D shows its transformed rank matrix E.

Since this transformation would allow our algorithm to only find over-expressed genes (genes with an average high rank within a subset of the samples), we duplicate each row and also assign ranks in the reverse order. This will allow to find under-expressed genes as well. For example, given an expression vector $g = (-2.0, -3.0, 2.0, 3.0)$, we obtain both rank vectors $r_o = (2, 1, 3, 4)$ (assigning high ranks to over-expressed genes) and $r_u = (3, 4, 2, 1)$ (assigning high ranks to under-expressed genes). As a consequence, the resulting rank matrix has twice as many rows as the original matrix. For ease of exposition we will consider matrix E to have the same size as A, but the algorithm trivially works on the duplicated matrix and we will use this extended version in the experiments.

Transforming mutation data and interaction network To transform the Boolean mutation matrix into a rank matrix, we first map each sample's mutation profile to the given interaction network (Vanunu *et al.*, 2010; Hofree *et al.*, 2013) and apply a network diffusion model. The obtained diffusion values are then transformed to ranks, so that higher ranks indicate that a gene is relatively "close" to a mutated gene (for a particular sample). Figure 1E illustrates this procedure.

That is, let $M \in \{0, 1\}^{m \times n}$ be the mutation matrix, where m is the number of mutation genes and n is the number of samples (as before), and let $G = (V, E)$ be the interaction network. Applying diffusion (Hofree *et al.*, 2013) to the n columns of matrix M using G results in a diffusion matrix $B \in \mathbb{R}^{m \times n}$. Finally, by ranking the rows for each column we obtain the *ranked diffusion matrix* $D \in \sigma_2^{m \times n}$, $\sigma_2 = \{1, \ldots, m\}$, which we use as input for the next step of the analysis.

2.2 Mining subtypes using rank matrix factorisation

The matrix factorisation model that we introduce aims to jointly *factorise* the two transformed rank matrices D and E into a set of k ranked factors, where k is an integer given by the user. One *factor* consists of a set of mutation genes, a set of expressed genes and a set of related samples. To provide some intuition for our approach, we first present an optimisation model for a single ranked factor. We then generalise this to k ranked factors using rank matrix factorisation.

Mining a single ranked factor As mentioned above, a *ranked factor* represents a group of samples that are consistently over/under-expressed in a subset of expression genes and that share the same affected genes in the ranked diffusion matrix.

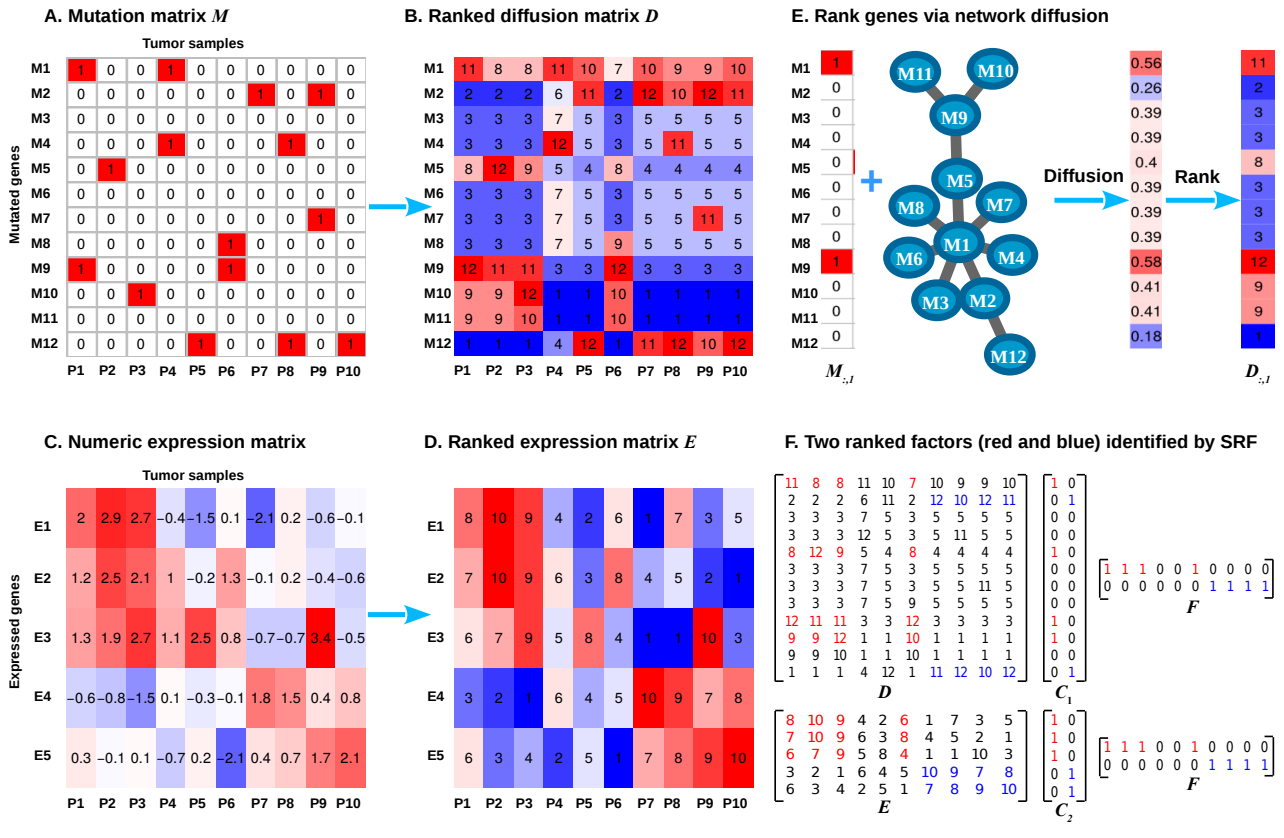

Fig. 1. SRF illustration. A) Boolean mutation matrix; B) Ranked diffusion matrix derived from the mutation matrix using the network diffusion model shown in Figure E and the parameter $\alpha = 0.7$; C) Numeric expression matrix; D) Ranked expression matrix, obtained by ranking column values in each row; E) Illustration how to derive a ranked diffusion vector for tumor sample $P1$ using his/her mutation profile and a given interaction network. F) Two ranked factors, represented by C_1, C_2 and F, identified by SRF in matrices D and E.

Let $\mathcal{P} = \{1, \ldots, n\}$, $\mathcal{M} = \{1, \ldots, m\}$ and $\mathcal{E} = \{1, \ldots, l\}$ be index sets for tumor samples, mutation genes and expression genes respectively. A ranked factor is represented by a tuple (P, G^M, G^E), where $P \subseteq \mathcal{P}$, $G^M \subseteq \mathcal{M}$ and $G^E \subseteq \mathcal{E}$. Inspired by our *ranked tiling* work (Le Van *et al.*, 2014), a ranked factor is obtained by optimising:

$$\begin{aligned} \text{argmax}_{P,G^M,G^E} \quad & \textstyle\sum_{m \in G^M, p \in P}(D_{m,p} - \theta_1) + \\ & \beta \textstyle\sum_{e \in G^E, p \in P}(E_{e,p} - \theta_2) \end{aligned} \tag{1}$$

subject to

$$\forall m \in \mathcal{M} : m \in G^M \quad \rightarrow \quad \textstyle\sum_{p \in P} M_{m,p} \geq \mu, \tag{2}$$

where θ_1 and θ_2 are user-defined thresholds that control how high ranks in D and E respectively should be to be included in the solution. We sometimes indicate these thresholds using relative values, i.e., $\theta_1 = a\%$; in this case, the absolute threshold is $\theta_1 = a\% * n$. β is a user-defined threshold to balance the contributions from the values in the two matrices. μ indicates the number of patients in which a mutation should be present in order to be included in the factor.

The objective in Equation 1 selects those rows (mutation and expression genes) and columns (samples) that together maximise the total sum of the values in the corresponding cells in the matrices, adjusted by θ_1 and θ_2. That is, cells that are lower than the thresholds are penalised and those that have higher values are rewarded. Equation 2 ensures that each gene that is selected from the ranked diffusion matrix is mutated in at least one of the samples present in the selection P. That is, genes that receive a high rank because they are in the network neighbourhood of genes mutated in the sample but are never mutated themselves will not be selected.

For example, given the matrices in Figure 1A, 1B, and 1D, and parameters $\theta_1 = 7, \theta_2 = 5, \beta = 1$, solving the objective results in $P = \{P1, P2, P3, P6\}$, $G^E = \{E1, E2, E3\}$, $G^M = \{M1, M5, M9, M10\}$. It is clear from the input matrices that this solution corresponds to an area with relatively high ranks. No more samples or genes can be added to the solution without decreasing the score. Note that mutated gene $M11$ was not selected despite having a high rank as no samples in the group carry a mutation for this highly ranked gene.

Mining k ranked factors using RMF As Equations (1) and (2) only provide a way to find a single factor, we here present a variation of rank matrix factorisation (Le Van *et al.*, 2015) to find a set of k non-redundant rank factors. To understand this generalisation, let us first reformulate the problem of finding one factor. Let us note that we can represent the set P using a $1 \times n$ boolean matrix F, with $F \in \{0, 1\}^{1 \times n}$, such that $F_{1,p} = 1$ iff $p \in P$. Similarly, we can represent the set G^M using an $m \times 1$ Boolean matrix C_1, such that $(C_1)_{r,1} = 1$ iff $r \in G^M$, and the set G^E using an $l \times 1$ Boolean matrix C_2, such that $(C_2)_{r,1} = 1$ iff $r \in G^E$.

Let us now denote by $J_{m,n}$ the $m \times n$ matrix in which all cells are filled with ones, and let the distance between two matrices be defined by $d(A, B) = \sum_{ij} A_{i,j} B_{i,j}$. Then we can rewrite our earlier problem as follows in matrix notation:

$$\arg \max_{C_1, C_2, F} \quad d(D - \theta_1 J_{m,n}, C_1 \odot F) + \beta d(E - \theta_2 J_{l,n}, C_2 \odot F) \tag{3}$$

$$\text{subject to} \quad \mu C_1 \leq M F^T. \tag{4}$$

Here, $\odot$ denotes the Boolean matrix product; in the Boolean matrix product the or operator is used instead of the addition operator, i.e., it is assumed that $1+1=1$. The constraint ensures that every selected mutated gene is present in the required number of patients; it uses the traditional matrix product. This notation makes the connection between our earlier problem and matrix factorisation clear: we are factorising the ranked data in a number of Boolean matrices that indicate where values of high rank can be found.

This new formulation is trivially extended to an arbitrary number of factors: by allowing that $C_1 \in \{0,1\}^{m\times k}$, $C_2 \in \{0,1\}^{s\times k}$, $F \in \{0,1\}^{k\times n}$ for arbitrary $k \geq 1$, we obtain a generic matrix factorisation setting for any k. Intuitively, in this setting we identify a number of rectangles in the data; the union of these rectangles is required to contain the highest ranks. An example is provided in Figure 1F.

To solve the optimisation problem, we follow the algorithm proposed by Le Van *et al.* (2015). That is, the SRF algorithm follows an iterative EM-style scheme, in which first C_1 and C_2 are optimised given F, and then F is optimised given C_1 and C_2. We repeat this iterative scheme until the optimisation score cannot be improved any further. When either C_1 and C_2 or F is known, it can be shown that the optimisation problem (3) – (4) is an integer linear programming (ILP) problem. Each such optimisation problem can be solved optimally. To avoid local maxima, we initialise the algorithm with a matrix F obtained by performing hierarchical clustering to cluster the columns into k groups.

As additional contribution, we develop a *parallel* implementation of the above algorithm, which makes it scalable to large datasets. Observe that Equations (3) – (4) allow each row of C_1 and C_2 to be optimised independently given F. Further, given C_1 and C_2, each column of F can be optimised independently if we relax the inequality in Equation (4), which puts a constraint on the columns and hence makes them dependent. However, if we require the iterative process to terminate after the step optimising C_1 and C_2 given matrix F, we still obtain a very good approximation upon convergence of the algorithm.

We implemented SRF in OscaR (OscaR Team, 2012) and used Gurobi as the back-end solver. The implementation is available at: `https://github.com/rankmatrixfactorisation/SRF`.

2.3 Deriving cancer subtypes from ranked factors

Ranked factors model groups of tumor samples that are homogeneous in gene expression as well as in mutations. Hence, if we obtain k *non-overlapping ranked factors*, i.e., factors that cover fully disjoint sets of samples, each factor found is considered to represent a unique subtype.

If the factors *overlap* in the sample dimension, however, we consider each group of samples that is covered by a unique combination of ranked factors to form a subtype. The reason for this is that each combination of ranked factors represents a different combination of expression and mutation profiles. In this case, the mutation and expression gene sets of a subtype are formed by the union of the mutation and expressed genes (respectively) of all factors in the combination. Section 3 shows examples of this concept. In practice, we prune subtypes covering fewer samples than a user-defined threshold, to avoid the discovery of small 'subtypes' that are most likely artefacts of noise in the data.

3 Results

3.1 Analysis overview

As input data we use 1) a gene–patient expression matrix describing for each patient its expression phenotype; 2) a gene–patient mutation matrix that describes per patient which of the genes carry somatic mutations (see Figure 1 for an overview of the analysis). To search for pathway level consistency across the tumor samples we use a transformed mutation matrix obtained by diffusing, per patient, the effect of each mutation over a given interaction network. In this way, not only genes that are mutated will receive high relevance scores, but also genes that are close to the mutated genes in the network. Identifying groups of patients with a consistent mutation profile in this transformed matrix allows searching for mutational consistency at the pathway level (Hofree *et al.*, 2013) and accounts for the clonality of carcinogenesis. Transforming the expression and mutation matrix to rank matrices is key to removing the scale differences.

A subtype is subsequently defined as a set of tumor samples that share a similar molecular origin of their disease, i.e., a driver pathway where the driver mutations occur. The effect of a mutated driver pathway is assumed to be reflected in the expression phenotype, consistently down- or upregulated compared to the reference, of a subset of the genes downstream in those samples. Hence, selected genes in the expression data and selected mutations in the mutation data of the samples in a subtype can be different.

Detecting a subtype is formalised as a complex bi-clustering problem in which one wants to search for a subset of patients that share both a similar set of driver mutations and a subset of consistently differentially expressed genes; given that the clonal phenotypes in cancer are affected in the same driver pathways, this assumption is reasonable. This clustering problem is solved by applying *rank matrix factorisation* (Le Van *et al.*, 2015) to jointly factorise the ranked mutation and expression matrices into a number of *ranked factors*. Conceptually, each resulting factor consists of a subset of samples associated to a subset of expression and mutation genes (expression and mutational features) for which the selected samples have respectively highly ranked expression values and highly ranked relevance scores. Expression and mutational features can, but do not have to overlap. Whereas a factor represents a group of patients together with their characteristic features, a subtype is defined as a group of patients covered by a unique combination of factors. Subtypes can thus mutually overlap in their characteristic expression and/or mutational features. This overlap in features between subtypes reflects the fact that subtypes are rarely distinct but rather represent a continuum of possible alterations. The subtyping algorithm is dubbed SRF, for *Subtyping with Ranked Factors*.

3.2 Results on simulated datasets

To test the performance of the method in recovering known subtypes, we generated datasets in which each subtype was defined as a set of tumor samples carrying a number of driver genes and a concomitant set of consistently over- and/or under-expressed genes of which the expression phenotype is assumed to be triggered by the driver mutations. The data contained 4 subtypes that occasionally shared genes mutated in the same driver pathways or genes displaying the same consistent expression. We imposed the rule that whenever two subtypes share genes mutated in the same driver pathway(s), they should share a set of consistently expressed genes. Figures 2A and 2B show an example dataset.

Driver genes were modeled to display mutational consistency at the pathway level across tumor samples belonging to the same subtype by selecting the drivers of those patients from a pre-selected set of genes that are *closely connected* in a real protein-protein interaction network and therefore assumed to belong to the same driver pathway. We varied the size of the driver pathways as well as the mutational recurrency of the driver genes for the samples within the subtypes to generate datasets (see Section 4).

For each simulated dataset, we ran our algorithm with varying parameter settings. We used two parameters to specify the preferred ranges of the ranks from the two input matrices, and another to balance the contributions of the two matrices. For each parameter setting, we used SRF to search for $k = 5$ subtypes, where the 5th subtype serves as the collection of tumor samples that have no clear subtype assignment. We initialised the

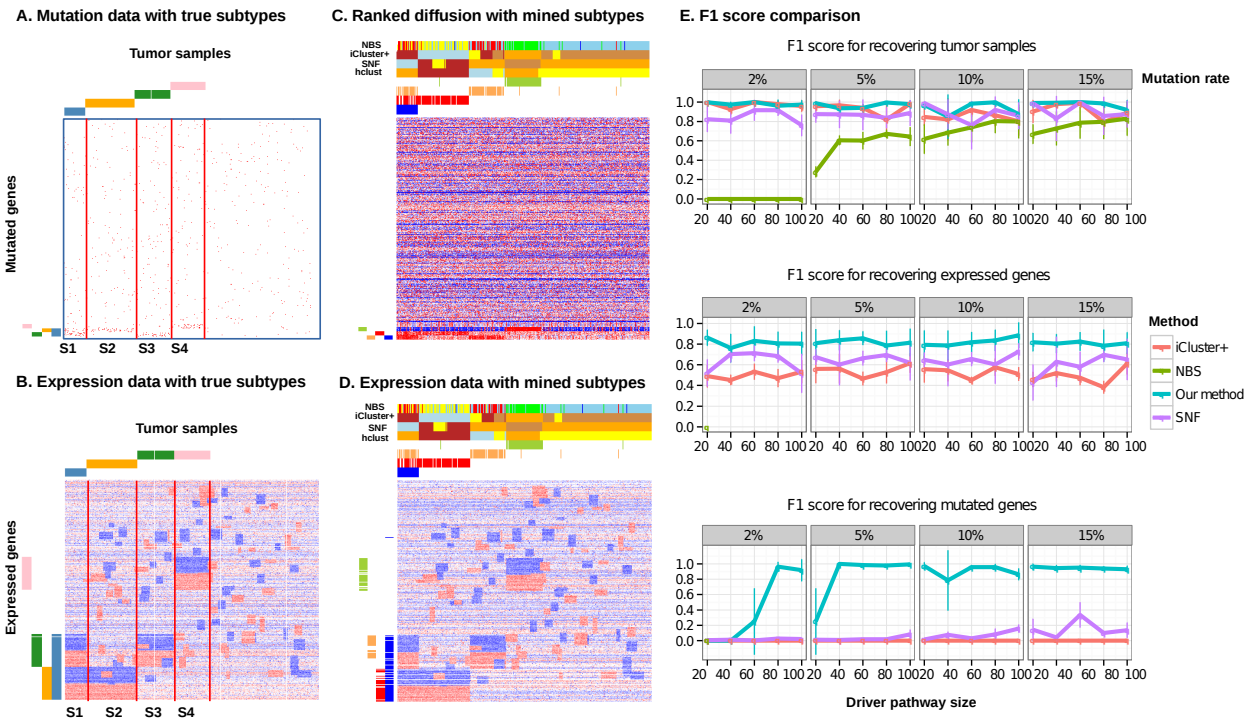

Fig. 2. Evaluation on simulated datasets. Panel A–B: Example data with ground truth. The heatmaps show mutation and numeric expression data of a representative simulated dataset, with a 10% mutational recurrency (meaning that a gene is mutated in at least 10% of the samples in a given subtype) and pathway size of 40. The four ground truth subtypes are marked by the horizontal and vertical coloured bars above and to the left of the heatmaps. Panel C – D: Results on the data shown in panels A and B. Results obtained by NBS (Hofree et al., 2013), iCluster+ (Mo et al., 2013), SNF (Wang et al., 2014) and the hierarchical clustering algorithm (hclust), which we used to initialize our model, are shown in the coloured bars above the heatmaps. The results obtained with SRF are indicated by the four coloured horizontal and vertical bars, just above and to left of the heatmaps; each bar indicates the patients (horizontally) and genes (vertically) selected by a ranked factor. Panel E: Performance comparison. The three plots denote F1 scores for 1) patient recovery (top), 2) expression gene recovery (middle), and 3) mutation gene recovery (bottom) for iCluster+, NBS, SRF, and SNF, for simulated datasets of varying driver pathway sizes and mutational recurrencies. Note that NBS does not work with expression data and we were unable to recover the mutated genes due to a lack of documentation.

algorithm with five sample groups obtained by a hierarchical clustering of the tumors using the ranked expression data.

After factorising the rank matrices, resulting subtypes containing less than 4% of the total number of samples were pruned (see Section 4).

Accuracy of the identified subtypes We evaluated the performance of our algorithm in recovering the known subtypes as well as their characteristic expression and mutational features. For this we used the F1 score, which assesses the trade-off between correctly and comprehensively distinguishing between samples, expression genes, and mutational features that truly belong to the subtypes from those that do not.

To optimise the parameter settings, we calculated F1 scores for different parameter settings and chose the one that resulted in the highest average score (see Section 4). Then, we used that parameter setting to evaluate the performance of the algorithm on all the simulated datasets.

Figure 2E shows the F1 scores obtained for different driver pathway sizes and mutational recurrencies. We can observe that the F1 score of recovering tumor samples of the simulated subtypes is high and largely independent of the sizes of the driver pathways and the mutational recurrencies. This demonstrates the added value of integrating the expression data. Further, the F1 scores of recovering mutation and expression genes relevant to the subtypes are generally high. As expected, the higher the mutational recurrency, the larger the number of mutated genes that can be recovered.

Comparison to related work To show that our method performs at least as well as state-of-the-art subtyping methods, we compared the results obtained by our method to those obtained with iCluster+ (Mo *et al.*, 2013), NBS (Hofree *et al.*, 2013) and SNF (Wang *et al.*, 2014), applied on the same simulated data. Both iCluster+ (Mo *et al.*, 2013) and SNF (Wang *et al.*, 2014) identify subtypes by jointly clustering the expression and untransformed mutation data, while NBS (Hofree *et al.*, 2013) exploits mutational information but does not use expression data.

Results obtained by iCluster+, NBS, SNF and SRF are summarised in Figure 2E. Our method obtained higher F1 scores than its competitors for both recovering expression and mutation genes. This is because our model couples genes, including mutation and expression genes, and patients to define subtypes and thus explicitly identifies subtype-specific genes. For iCluster+ and SNF, this is not the case and the selected expression or mutation genes are thus always the same, irrespective of the subtype.

As a representative example, we illustrate in more detail the results produced by the different methods on the simulated dataset with a 10% mutational recurrency and a pathway size of 40 (Figure 2C - 2D). The figures show that compared to the other methods, our method can discover overlap between very similar subtypes in both the patient and gene dimension. In addition, our model is shown to be tolerant to noise: the fifth ranked factor found by SRF remained empty, revealing that no 'noisy' patients and genes were incorrectly marked as belonging to a subtype.

3.3 Results on the TCGA breast cancer data

To test SRF in a real-world setting, we applied it to the well-studied TCGA breast cancer dataset. The method was ran as outlined in Section 4. As we were mostly interested in identifying subtype-specific features, we chose stringent parameters to only identify subtypes with representative profiles in terms of expression and mutations Factorising the dataset into $k = 8$ factors resulted in 13 subtypes. The number of identified subtypes is higher

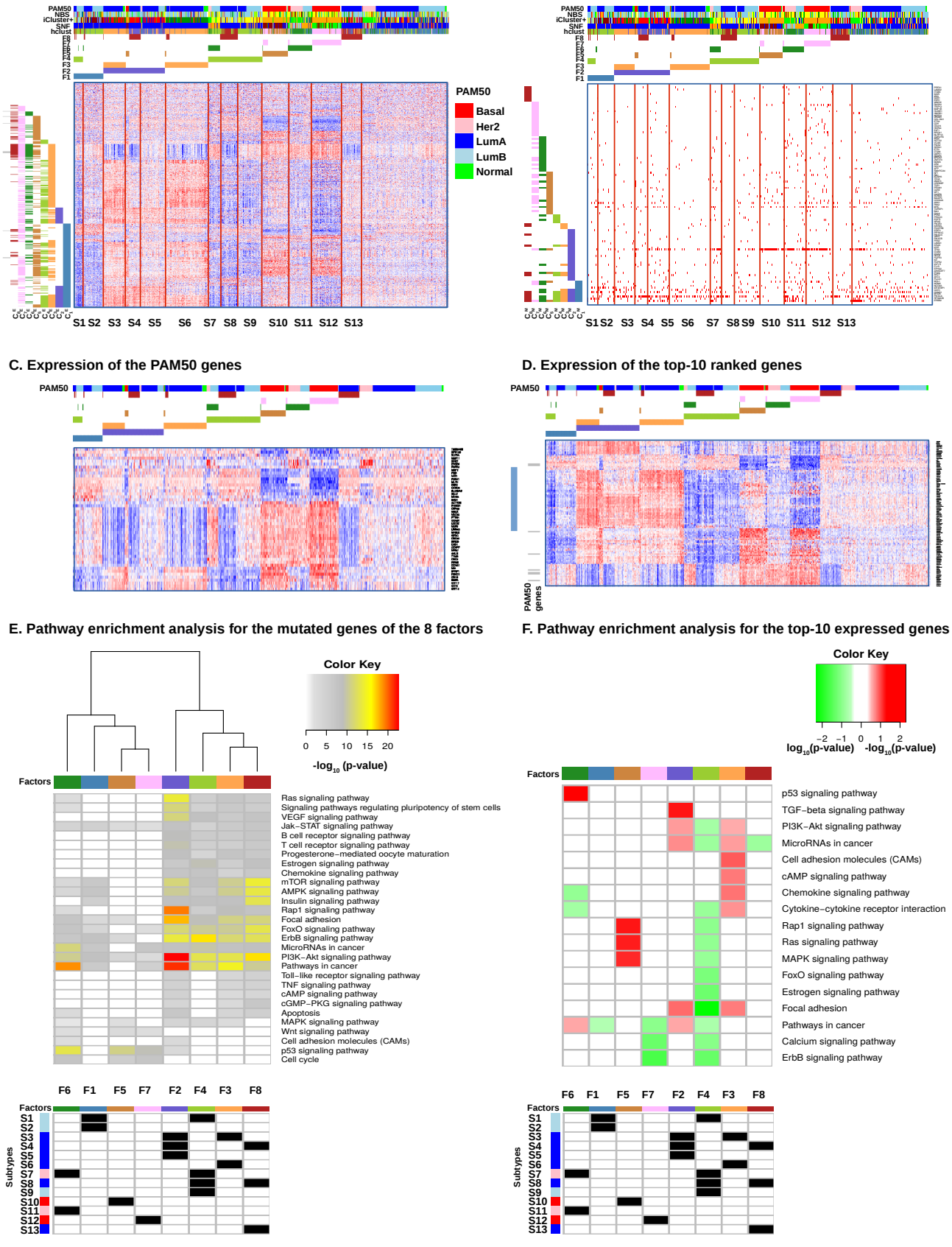

Fig. 3. Results of applying SRF ($k = 8$) on the TCGA breast cancer dataset, which resulted in the 13 subtypes denoted by S1, . . ., S13. For heatmaps in Panels A, C and D: red implies over-expressed, white neutral, blue under-expressed. Panel A: Expression data. The gene and tumor sample sets corresponding to the eight ranked factors are marked by the vertical and horizontal colour bars. Each subtype is a unique combination of ranked factors in the tumor sample dimension. Panel B: Mutation data. Samples are ordered as in panel A. Note that although the method uses the diffused ranked mutation matrix, the shown heatmap corresponds to the Boolean mutation data prior to diffusion. Only mutated genes that belong to any of the factors are displayed. On panels A and B, the top bar indicates the PAM50 annotation of the samples together with the subtyping results of iCluster+, NBS, SNF, and hierarchical clustering. Panel C: Expression heatmap of the PAM50 genes. The samples are ordered as in panel A. Panel D: Expression heatmap of the top-10 genes per subtype, i.e., the ten genes having the highest average ranked scores per subtype. Panel E: KEGG pathway (Kanehisa and Goto, 2000) enrichment results. Mutated genes of the eight factors were tested for pathway enrichment; resulting $-log_{10}$ p-values are shown for cancer related pathways that were found to be significantly enriched in at least one of the factors. Panel F: KEGG pathway enrichment analysis for the top-10 ranked expressed genes of the eight factors. log_{10} and $-log_{10}$ p-values are shown for pathways having under- and over-expressed genes respectively.

than the number of factors because subtypes are defined as combinations of factors (see Section 2.3). The results are visualised in Figure 3.

To validate our subtypes, we tested 1) to what extent the discovered subtypes corresponded to the PAM50 classification, and 2) to what extent SRF could further refine it. Figures 3 and 4A show that most subtypes are enriched in samples with the same PAM50 label (Parker *et al.*, 2009) as shown in Figure 3A. All samples of the same PAM50 class rarely end up in a single subtype. The Basal subtype, for example, is divided into two major subgroups: S10, S12; LumA is divided into S3, S4, S5, S6, S8, S13; LumB into S1, S2, S5, S9; Her2 into S11 and S7. So our approach does not only match the PAM50 classification to a large extent, it also further refines known subtypes.

This high-resolution subtype refinement is a characteristic property of the method's intrinsic feature selection. Rather than using global profiles to group samples, the methods actively searches for combinations of feature sets (factors) that characterise samples using rather stringent criteria. As a result differences between expression and mutational profiles are marginal for some subtypes (e.g., for LumA-related subtypes S3, S4, S8, S13, and for LumB-related subtypes S7 and S9). Retrospectively, it might have been possible to merge these subgroups. However, in case of subtypes S10 and S12, carrying samples with the same Basal label, the subtype-specific mutational and gene expression profiles are quite distinct for the selected feature sets, corresponding to the brown and pink bars in Figure 3.

Next to the subtypes that have rather homogeneous PAM50 labels assigned to their samples, subtypes S1, S2, S5 and S9 contain a mixture of LumA and LumB samples, and S11 contains a mixture of Her2 and LumB samples. Although some inconsistency between the mere expression-based PAM50 classification and subtyping protocols based on the integration of expression and genomic information is to be expected (Curtis *et al.*, 2012), a closer inspection of the expression and mutational profiles of the subtypes with mixed PAM50 class membership shows why our method does not distinguish between, e.g., the selected Her2 and some LumB samples. That is, the selected LumB samples of subtype S11 contain clear Her2-related features that distinguish them from other LumB samples, such as an increased ERBB2 amplification and a more pronounced over-expression of a characteristic subset of genes.

Our approach towards identifying subtypes together with their features can only be meaningful if the selected features are biologically relevant. To assess this, we first tested to what extent the expression features used to build the PAM50 classifier are amongst our selected features. From the 50 PAM50 features, 49 were present amongst the features selected by our method after pre-processing (see Section 4). The ranked factors found by SRF used 2221 features in total, including 48 out of 50 PAM50 features. To select a smaller representative feature set, the 10 genes with the highest average ranked score per subtype were selected, resulting in 110 instead of 2221 features (Figure 3D). Those 110 features contained 8 out of the 48 remaining features of the PAM50 classifier. SRF selects all high-ranking features, hence the selected feature sets are more redundant than those used by PAM50, which were designed for classification. If our approach is to coincide with PAM50, we expect each group of features to be covered by a few PAM genes. Except for one subtype this is indeed the case. The exception, indicated with the blue row bar in Figure 3D, does not have a corresponding PAM50 feature. Remarkably this is the feature set that has the most distinct difference in expression between the two subtypes with the same PAM50 Basal label (S10 and S12). Figure 3C shows how indeed no differences can be observed between Basal subtypes S10 and S12 using the PAM50 features, whereas the subdivision is clear using the expression-based features selected by our method and shown in Figure 3D. Figure 5 shows how the subtype subdivision of the Basal-like subtypes and the selection of the corresponding expression features is also driven by the simultaneous selection of the mutation-based features: S10 and S12

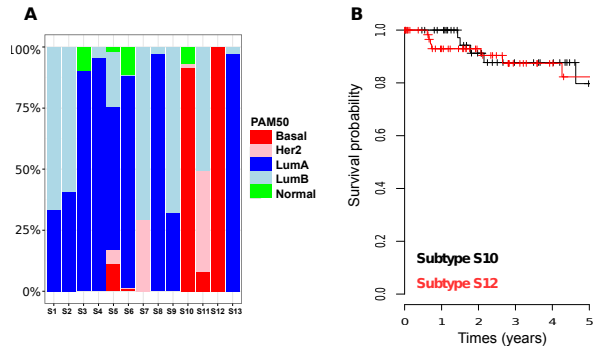

Fig. 4. A: Distribution of PAM50 samples in the identified subtypes; B: Kaplan-Meier plot for the two Basal-related subtypes S10 and S12. (Not statistically significant because of the low mortality rate.)

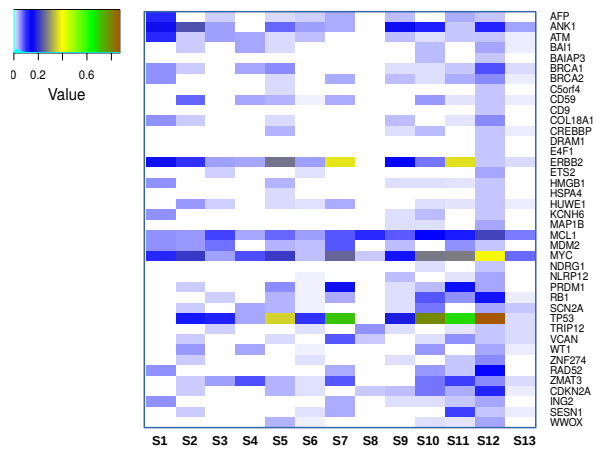

Fig. 5. Comparing the mutation frequencies among the identified subtypes using the mutation gene set selected by subtype S12.

show clearly distinctive mutational profiles with different mutational frequencies of, e.g., *CD9, DRAM1* and *E4F1*. Interestingly, survival analysis of these two Basal-like subtypes, despite not being significant due to the low mortality rate, shows that subtype S12 tends to be more aggressive than S10 during the first two years (Figure 4B).

To assess whether the selected feature sets belong to known driver pathways in breast cancer, we did per factor a KEGG pathway enrichment analysis on respectively the selected mutational and expression features. Figures 3E and 3F display the enrichment levels for a representative set of cancer related pathways that were found to be enriched. They also indicate how each subtype is a composition of different factors and how the factors overlap in genes and thus also in enriched pathways. For instance, the genes with a characteristic expression profile in factor 5 (representative for lumA and B) and factor 4 (representative for basal S10) are enriched in rap1, ras1 and mapK signaling, but with an anticorrelated expression profile for the Luminal subtypes versus the Basal one. Figure 3E shows how, as expected, (Cancer Genome Atlas Network, 2012; Toss and Cristofanilli, 2015; Verbeke *et al.*, 2015), the Basal-related (S10, S12) and Her2-related subtypes (S11) are highly enriched in *p53 signaling* and *cell cycle* whereas other subtypes are not. In addition, the Luminal subtypes (LumA and LumB; S2, S5, S6, S9 and S13) are enriched with cancer pathways known to be specific for this group: *PI3K-Akt signaling pathway* (Cancer Genome Atlas Network, 2012; Verbeke *et al.*, 2015), *Estrogen signaling pathway* (Lisa *et al.*, 2013), *AMPK signaling pathway* (Verbeke *et al.*, 2015).

Comparison with state-of-the-art methods

To test to what extent our method agrees with state-of-the-art subtyping methods, we also ran iCluster+ (Mo *et al.*, 2013), NBS (Hofree *et al.*, 2013) and SNF (Wang *et al.*, 2014) on the same dataset. Parameter settings for each of these methods was optimised as explained in Section 4. Figure 3 illustrates how our results compare with those of the other tools in terms of matching the PAM50 subtyping. SNF and iCluster+, the two integrative methods that do not use mutational consistency at the pathway level, do not perform well for some subtypes. For example, SNF could not discern the heterogeneity of the Luminal samples, which has been known to be the most heterogeneous breast cancer subtype (Cancer Genome Atlas Network, 2012; Curtis *et al.*, 2012), when it clustered all LumA samples and a large number of LumB samples into one group; iCluster+ could not subdivide the Basal subtype. NBS, which could use mutational consistency at the pathway level but could not integrate with expression data, could not distinguish the LumA from the LumB samples. In contrast, our method can integrate expression data and mutational data at the pathway level and hence can capture subtle differences that might be missed otherwise.

Comparison with hierarchical clustering We performed a hierarchical clustering of the tumor samples into clusters, of which the result is annotated by the *hclust* column bar in Figure 3A. This clustering is also used to initialise the matrix F of the factorisation in Equation (3). In contrast to hierarchical clustering, our SRF algorithm identifies clusters (subtypes) that are highly overlapping and removes noisy samples.

4 Materials and methods

Simulated data Mutational data was generated by first selecting driver pathways for each of the simulated subtypes. Driver pathways were selected from the densely connected sub-networks obtained by applying the *InfoMap* algorithm (Rosvall and Bergstrom, 2008) implemented in the *igraph* (Csardi and Nepusz, 2006) R package on the STRING network (Szklarczyk *et al.*, 2011) post-processed by Hofree *et al.* (2013). The selected driver pathway sizes varied from $\{20, \ldots, 100\}$ genes and each such gene was assigned a mutational recurrency between 2% and 15%. Passenger mutations were simulated by sampling, for each patient, from a Bernoulli distribution with $p = 0.005$ (the average mutational recurrency we observed in the TCGA breast cancer data). The total number of passengers was chosen such that the total number of mutation genes, including both drivers and passengers, was 8000, which was approximately equal to the number of mutation genes used in the TCGA breast cancer data. Each simulated mutation matrix consists of 8000 genes $\times$ 350 patients.

Expression data was simulated as previously described by Le Van *et al.* (2015). That is, first background information was generated by sampling from a mixture of three Gaussians, of which means were uniformly sampled from three different ranges, namely, [-5,3), [-3,3] and (3,5]. Then, over-expressed and under-expressed modules were implanted. Values within over-expressed and under-expressed modules were sampled from a Gaussian, with mean uniformly sampled from (3,5] and [-5,-3) respectively. Per set of driver mutations and thus per subtype, we ensured that the simulated dataset consisted of at least one set of genes that was consistently differentially expressed across the samples in the subtype.

To simulate noise in the expression data, we simulated 100 small expression modules that could be seen as the result of some confounding factors such as sex and tissue type. The number of rows and columns of these confounding modules were sampled from a normal distribution, whose mean was equal to 25% of the medium-sized pattern and standard deviation was equal to 40% of the mean. Each simulated expression matrix consists of 4000 genes $\times$ 350 patients.

TCGA breast cancer dataset Breast cancer somatic mutation, copy number alteration (CNA), expression (RNA-Seq v2), and clinical data were downloaded from the TCGA data portal. Mutational data were converted to a Boolean mutation matrix. CNA data were analysed using Gistic 2.0 (Mermel *et al.*, 2011) with default settings. This data was then binarised by considering how genes are classified by Gistic: as either deleted or amplified. This information was added to the mutation matrix. We restricted our analysis to mutations and CNVs in genes that also appear in the STRING network (12232 vertices) prepared by Hofree *et al.* (2013). Expression genes were selected based on their differential expression relative to normal (non-tumor) samples: for each gene a normal distribution was fitted using the normal expression samples and z-scores were calculated for the tumor samples. We then evaluated the 5th and 95th percentiles of the tumor samples. Genes were selected if 1) the p-values for these percentiles were below 0.001 and 2) their log-fold change relative to the mean normal expression was at least 2.5. After the filtering steps mentioned above, the final mutation matrix consisted of 8604 genes $\times 719$ patients, and the final expression matrix of 2472 genes $\times$ 719 patients.

Parameter selection To allow for a fair comparison, parameters of the algorithms were optimised to obtain the best possible results. See the supplementary document for a more detailed discussion.

Gene feature selection To compare SRF to other methods concerning the recovery of subtype-specific genes in the simulated datasets, we used gene feature selection. With our method, it was straightforward to extract the mutation genes and expression genes representative of the individual subtypes, as described in Section 2. With iCluster+ (Mo *et al.*, 2013), we used a quantile cut-off of $p = 0.75$ to select the important genes according to the model. With SNF (Wang *et al.*, 2014), we first ordered the genes by *Normalized Mutual Information* using the SNF software. We then selected the top-n expression and the top-m mutation genes, where n and m are the total number of true expression and mutation genes of the simulated subtypes respectively. With NBS (Hofree *et al.*, 2013), we could in theory obtain subtype-specific mutation genes, but were not able to recover them given the lack of documentation. It is important to note that with both iCluster+ and SNF, all identified subtypes have the same set of mutation and expression genes.

Hierarchical clustering SRF requires an initialised matrix F to start from, which was obtained through hierarchical clustering: we used the *hclust* package in R to cluster the columns of the ranked expression matrix into k groups (with Euclidean distance).

Pruning small subtypes To be more tolerant towards noise, derived subtypes that contain less samples than a predefined threshold (less than 4% of the total number of samples for the simulated datasets) were pruned. Samples of the pruned subtypes were re-assigned to the remaining subtype that results in the highest score for the function in Equation 1.

Survival and pathway enrichment analysis Survival analysis was performed using the R *survival* package. We used time to follow, time to event and the subtype information produced by our algorithm to calculate the survival probability. Pathway enrichment analysis was done using the ClueGo plugin (Bindea *et al.*, 2009) in Cytoscape (Shannon *et al.*, 2003).

5 Discussion

Previous integrative models, such as iCluster+ (Mo *et al.*, 2013) and SNF (Wang *et al.*, 2014), used between-sample similarities from the sample's global expression/mutational profiles to derive subtypes. However, molecular subtypes are defined by the molecular mechanisms that drive carcinogenesis. How subtypes are defined thus depends on the features used to group samples in subtypes. Conversely, the subtypes define which features are relevant for a certain sample grouping. Hence, subtyping and feature identification are confounded problems that ideally should be solved simultaneously.

In this work we therefore developed SRF, an approach that does so. To this end we approached the subtyping problem by decomposing patient–mutation and patient–expression data into ranked factors. A factor here represents a set of samples for which a set of genes display mutational consistency and a (possibly overlapping) second set of genes display expression consistency. A factor thus is an expressed and mutational feature set shared by a group of samples, and can be viewed as a *bi-cluster* (Madeira and Oliveira, 2004) in respectively the expression and mutation data that are coupled in the patient dimension. We developed a *global model* in the form of matrix factorisation to identify these bi-clusters.

Subtypes are then defined as each patient set that is covered by a unique combination of ranked factors. As a result subtypes can overlap in the factors that characterise them, reflecting the fact that subtypes are never mechanistically completely different, but share common representative features/driver mechanisms.

Compared to state-of-the-art methods, our method is most related to Hofree *et al.* (2013) 1) as it uses a network model to account for pathway level parallelism between independently evolved tumor samples and 2) because it extracts features explicitly. However, it is different from Hofree *et al.* (2013) by integrating both mutational and expression data.

Compared to related methods, samples without clear-cut signals will not be assigned to any subtypes. This prevents samples (noisy or heterogeneous samples) from blurring the molecular characteristics that are representative for a subtype. However, if desired we could assign samples to the closest subtype identified by our method.

In this paper, we did not consider metabolomics data, as was done in the work by Tardito *et al.* (2015). The type of data described in Tardito *et al.* (2015) can only be generated for cell lines and not for a tumor biopsy. It is not available in the context of TCGA or ICGC.

We extensively tested the performance of our method on simulated data. Testing and comparing our method with other state-of-the-art subtyping methods on the well studied TCGA breast cancer dataset shows how our method is able to grasp the most prominent signatures in the data. In addition, however, our method is also able to capture subtle differences that are missed by methods that compare samples based on global profiles of similarities.

Acknowledgements

The authors acknowledge support from Ghent University Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"; Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) [G.0329.09, 3G042813, G.0A53.15N]; Agentschap voor Innovatie door Wetenschap en Technologie (IWT) [NEMOA and the personal fellowship of Dries de Maeyer]; Katholieke Universiteit Leuven [PF/10/010] (NATAR); the Research Foundation–Flanders (FWO) project "Instant Interactive Data Exploration".

References

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., *et al.* (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**(8), 1091–1093.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61–70.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.

De Maeyer, D., Weytjens, B., De Raedt, L., and Marchal, K. (2016). Network-based analysis of eqtl data to prioritize driver mutations. *Genome Biology and Evolution*.

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature methods*, **10**(11), 1108–15.

Kanehisa, M. and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**(1), 27–30.

Le Van, T., van Leeuwen, M., Nijssen, S., Fierro, A. C., Marchal, K., and De Raedt, L. (2014). Ranked tiling. In *ECML PKDD 2014 (2)*, pages 98–113.

Le Van, T., van Leeuwen, M., Nijssen, S., and De Raedt, L. (2015). Rank matrix factorisation. In *PAKDD 2015*, pages 734–746. Springer.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., *et al.* (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, **47**(2), 106–114.

Lisa, B., Friederike, M., Petra, S., Andreas, S., *et al.* (2013). Intrinsic breast cancer subtypes defined by estrogen receptor signalling - prognostic relevance of progesterone receptor loss. *Mod Pathol*, **26**(9), 1161–1171.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, **1**(1), 24–45.

Mermel, C., Schumacher, S., Hill, B., Meyerson, M., *et al.* (2011). Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, **12**(4), R41.

Mischel, P. S., Shai, R., Shi, T., *et al.* (2003). Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene*, **22**(15), 2361–73.

Mo, Q., Wang, S., Seshan, V. E., Olshen, *et al.* (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *PNAS*, **110**(11), 4245–50.

OscaR Team (2012). OscaR: Scala in OR. Available from `https://bitbucket.org/oscarlib/oscar`.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., *et al.* (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, **27**(8), 1160–7.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., *et al.* (2000). Molecular portraits of human breast tumours. *Nature*, **406**(6797), 747–52.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *PNAS*, **105**(4), 1118–1123.

Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., *et al.* (2014). Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. *Cell*, **159**(6), 1461–1475.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498–2504.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, **98**(19), 10869–74.

Speicher, N. K. and Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, **31**(12), i268–i275.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., *et al.* (2011). The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, **39**(suppl 1), D561–D568.

Tardito, S., Oudin, A., Ahmed, S. U., Fack, F., *et al.* (2015). Glutamine synthetase activity fuels nucleotide biosynthesis and supports growth of glutamine-restricted glioblastoma. *Nat Cell Biol*, **17**(12), 1556–1568.

Toss, A. and Cristofanilli, M. (2015). Molecular characterization and targeted therapeutic approaches in breast cancer. *Breast cancer research*, **17**(1), 60.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., *et al.* (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, **14**(16), 5198–5208.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, **6**(1), e1000641.

Verbeke, L. P. C., Van den Eynden, J., Fierro, A. C., Demeester, P., Fostier, J., and Marchal, K. (2015). Pathway relevance ranking for tumor samples through network-based data integration. *PLoS ONE*, **10**(7), 1–22.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., *et al.* (2013). Cancer genome landscapes. *Science*, **339**(6127), 1546–1558.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**(3), 333–7.

Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol*, **7**(10), e1002227.

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, *et al.* (2013). Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, **45**(10), 1134–1140.